The 1st Temporal Web Analytics Workshop (TWAW)

Ricardo Baeza-Yates
Yahoo! Research
Barcelona
Spain
rby@yahoo-inc.com

Julien Masanès
Internet Memory Foundation
Paris
France
julien@internetmemory.org

Marc Spaniol
Max-Planck-Institut für Informatik
Saarbrücken
Germany
mspaniol@mpi-inf.mpg.de

ABSTRACT
The objective of the 1st Temporal Web Analytics Workshop (TWAW) is to provide a venue for researchers of all domains (IE/IR, Web mining etc.) where the temporal dimension opens up an entirely new range of challenges and possibilities. The workshop's ambition is to help shaping a community of interest on the research challenges and possibilities resulting from the introduction of the time dimension in Web analysis. The maturity of the Web, the emergence of large scale repositories of Web material, makes this very timely and a growing sets of research and services (recorded future[1], truthy[2] launched just in the last months) are emerging that have this focus in common. Having a dedicated workshop will help, we believe, to take a rich and cross-domain approach to this new research challenge with a strong focus on the temporal dimension.

Categories and Subject Descriptors
H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms
Algorithms, Management, Measurement, Documentation, Experimentation

Keywords
Temporal Web Analytics, Web Scale Data Analytics, Distributed Data Analytics

1. INTRODUCTION

Marking the World Wide Web's 20th and the Internet Archive's 15th anniversary in 2011, we now have a large collection digitally-born content from almost two decades. These archives not only capture the history of born-digital content but also reflect the zeitgeist of different time periods over more than a decade. This is already and will become more and more a gold mine for researchers, for instance sociologists, political scientists, media and market analysts, as well as experts on intellectual property (IP, e.g., at patent offices) etc.

Research on the temporal dimension of Web contents opens up great opportunities for analysts. For example, one could compare the notions of "online friends" and "social networks" as of today versus five or ten years back. Similar examples relevant for a business analyst or technology journalist could be about "tablet PC" or "online music". Similarly, the hyperlink structure of archived material can now be systematically exploited. It makes it possible to see how site (or even domain) structures develop over time, whether they are affected by Web spam or not, and which prevalent structures exist in general or within a certain domain.

The focus of TWAW and the topics addressed are a "natural" match with the WWW conference. With the World Wide Web celebrating its 20th anniversary in 2011 the need for a more systematic exploitation of our digital cultural heritage becomes evident. While the early 90's of the Web have been almost completely lost, national libraries, digital news archives and archiving institutions (like the Internet Archive Foundation) have protected Web contents from vanishing. These data are a potential goldmine for temporal Web analytics at the Web scale content level. However, the societal as well as scientific impact of temporal Web analytics has been insufficiently studied so far. As the WWW conference is the premier event series in this domain, we consider TWAW an ideal venue to exchange knowledge about temporal analytics on the Web scale with experts from science and industry.

2. WORKSHOP TOPICS AND THEMES

TWAW focuses on investigating infrastructures, scalable methods, and innovative software for aggregating, querying, and analyzing heterogeneous data at Internet scale. Particular emphasis will be given to temporal data analysis along the time dimension for Web data that has been collected over extended time periods. A major challenge in this regard is the sheer size of the data it exposes and the ability to make sense of it in a useful and meaningful manner for its users. It is worth noting that this trend of using big data to make inferences is not specific to Web content analytics. A now-common strategy in post-genomic biology is to measure, quantitatively, the action of all (or as many as possible) of the genes at the level of the transcriptome, proteome, metabolome and phenotype, and to use computerised methods to infer gene function via various kinds of pattern

[1] https://www.recordedfuture.com/

[2] http://truthy.indiana.edu/

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

recognition techniques. On the Web, we have to a large extent, also reached this point. Web scale data analytics therefore needs to develop infrastructures and extended analytical tools to make sense of these. Workshop topics of TWAW therefore include, but are not limited to following:

- Web scale data analytics
- Temporal Web analytics
- Distributed data analytics
- Web science
- Web dynamics
- Data quality metrics
- Web spam
- Knowledge evolution on the Web
- Systematic exploitation of Web archives
- Large scale data storage
- Large scale data processing
- Data aggregation
- Web trends
- Topic mining
- Terminology evolution
- Community detection and evolution

3. ORGANIZATION

The workshop is the first of its kind. Covering this novel and challenging research area of temporal Web analytics, the workshop organizers teamed up from an archiving institution, industry and research. Similarly, the international program committee consists of well renowned experts in one or more of topics addressed. The program committee consisted of:

- Eytan Adar (University of Michigan, USA)
- Omar Alonso (Microsoft Bing, USA)
- Srikanta Bedathur (IIIT-Delhi, India)
- Andras Benczur (Hungarian Academy of Science)
- Klaus Berberich (Max Planck Institute for Informatics, Germany)
- Adam Jatowt (Kyoto University, Japan)
- Scott Kirkpatrick (Hebrew University Jerusalem, Israel)
- Christian König (Microsoft Research, USA)
- Frank McCown (Harding University, USA)
- Michael Matthews (Yahoo! Research, Barcelona)
- Kjetil Norvag (Norwegian University of Science and Technology, Norway)
- Thomas Risse (L3S Research Center, Germany)
- Pierre Senellart (Télécom ParisTech, France)
- Torsten Suel (NYU Polytechnic, USA)
- Masashi Toyoda (Tokyo University, Japan)
- Peter Triantafillou (University of Patras, Greece)
- Gerhard Weikum (Max Planck Institute for Informatics, Germany)

4. ACKNOWLEDGMENTS

The organization of this workshop is partially supported by the 7th Framework IST programme of the European Union through the focused research project (STREP) on Longitudinal Analytics of Web Archive data (LAWA) under contract no. 258105 (cf. www.lawa-project.eu).